Available online at www.sciencedirect.com

ScienceDirect

Procedia Computer Science 87 (2016) 281 – 287

Procedia Computer Science

4th International Conference on Recent Trends in Computer Science & Engineering

Efficient Webpage Retrieval Using WEGA

Dr.J. Dafni Rose[a], J.Komala[b*], M.Krithiga[a,b]

[a]*Associate professor, St. Joseph's Institute of Technology, OMR, Chennai- 119, India*
[b]*Student, St. Joseph's Institute of Technology, OMR, Chennai- 119, India*
[a,b]*Student, St. Joseph's Institute of Technology, OMR, Chennai- 119, India*

Abstract

Web page retrieval plays a major role nowadays. The proposed system helps in the development of optimized web page recommendation. The proposed system is a novel method to provide the web pages effectively and accurately by using genetic algorithm. First the entered query is pre- processed and similar word set is generated. It uses word net tool for semantic keyword set generation that is the various fields where that particular word is used. This tool uses the ontology concept for the keyword generation. The user has to click a word from these keyset or directly click the pre- processed word. The selected keyword is given as input to server and some webpages will be retrieved. The web contents are extracted from these webpages. Then the weightage of that particular word is calculated in all the documents/web contents retrieved. Weight Enhanced genetic algorithm which involves three main process that is selection, crossover and mutation is applied to the weights calculated. Finally, the webpage with highest weightage will be displayed first and then consequently remaining webpages will be displayed in the same manner.

Keywords: *Weight Enhanced genetic Algorithm; Word net; Semantic Keyword Set; Ontology*

1 Introduction

Web page recommendation is one of the toughest and complicated jobs to be performed. Predicting the user's choice of webpage is the crucial point in webpage recommendation process. Many methods have been developed in the recent decades but almost all the techniques use web log files for predicting the user's choice of webpages. The user's navigational pattern is always used as a base for webpage recommendation. But not all the user needs webpages based on navigational patterns rather some user may need an unique webpages that other user has not yet visited. In order to overcome this issue, we suggest a new idea which concerns mainly about the relevancy of the input query. This can be done by using the Genetic Algorithm. Initially the weight of the keyword or pattern given as input query is calculated in each of the webpages. Then the Genetic Algorithm is applied to the weights calculated. This process can be explained in further sections of this paper.

This paper provides us with a better method for the web page recommendation issue. The proposed method does not concern about the navigational pattern of the users rather it focuses only on the relevancy of the input query given by the users. It also proves to be a better method compared to other webpage recommendation methods. This method has a greater advantage as it displays web pages based on the weights calculated but does not require any complex steps to be performed. It follows a simple yet effective method. In order to understand the concepts in a better way, the concepts are discussed in detail in various sections of these paper. The literature Review is discussed in section 2.The proposed system is discussed in section 3.The various steps followed are discussed in detail in this section. The result analysis is discussed in section 4 and the conclusion and future work is discussed in section 5.

2. Literature Survey

Many methods had been employed for webpage recommendation such as sequential modelling, Markov model and Tree based structure in the earlier period and it has shown some positive effect in the

Peer-review under responsibility of the Organizing Committee of ICRTCSE 2016
doi:10.1016/j.procs.2016.05.162

recommendation process.This paper is mainly concerned with predicting the user' s future request based on the evaluation done on past and current request of users [2][1].

Later Sematic Enhanced approaches has been applied to it. In those system Domain ontology helps in clustering the webpages, classifying them and in identifying the subject it belongs to. Many different algorithms has been used for webpage recommendation. Usually knowledge discovery from the available web usage data and providing a satisfactory knowledge representation is a very challenging process in web usage recommendation. Integrating the domain knowledge with the web usage knowledge in a the process of web page recommendation helps to provide a better results in case of semantic enhanced approaches. It provides a better results when compared to one of the web usage mining methods PLWAP- Mine[10].

Some new approaches have used traditional approaches by including additional parameters to it. The patterns include web usages data personalization, user navigational frequency analysis, session wise data analysis and time stamp data analysis. This method is mainly concerned with providing users with the rarely accessed patterns. This method uses the weight prediction method to identify the user pattern exactly. This method results in isolation of the user personalization, reduced complexity and less resource consumption. Some more additional parameter is needed to make the webpage recommendation more accurate and also to reduce the training time[6][7].

Searching the document from a huge set of document collection is very challenging and complicated job. Genetic Algorithm is very efficient in searching from a large set in a huge search spaces[11]. It helps in providing an accurate results given a huge search spaces. It provides greater accuracy and provide more optimization with respect to the change in population size. It helps to provide reduced search time[3][4].

3. Proposed System

In this paper, the objective is to develop a webpage recommender system that provides the exact search that is requested by the user. The webpage recommendation that involves now may deliver some irrelevant webpages and ads to the user. To find the exact webpage or result they need, may take few more minutes. In order to overcome this challenge, we are proposing a new system known as WEGA (i.e. Weight Enhanced Genetic Algorithm). The proposed system will provide exact result to the user as the web contents are processed to fulfil the user' s request. Then additionally this paper involves generation of keysets that predicts the various aspects of that pre- processed word. So that the user also the word from that keysets for recommending the webpages.

Weight Enhanced Genetic Algorithm uses the basic process of Genetic Algorithm. This model will takes less time, memory and resources. This satisfies user' s request and provide better results. This technique analyse the user' s request very deeply and deliver the webpages more accurately. The webpages displayed will be very effective.

3.1 System Architecture:

The mechanism of providing the exact web results to user is described using fig 1.

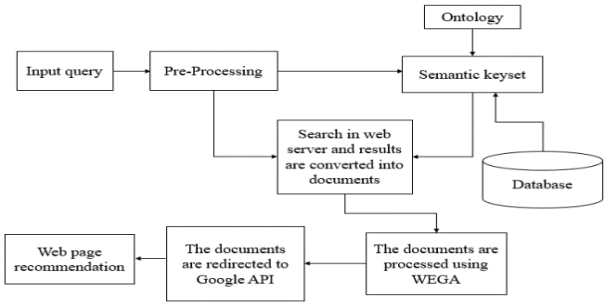

Fig 1. System Architecture

3.2 Input Query

The user has to login/ register to enter the search space. If the user is a new user, then registration process has to be carried out first and then the login process has to be followed. If the user is the old user, login process has to be done directly. After user login process, web user can enter the search space page.

3.3 Pre- Processing

Pre- processing method plays a very important role in text mining techniques and applications. It is the first step in the text mining process. It involves stop word removal and stemming process.For example, if the query is " How to add tiles to start menu in windows 10" then the pre- processed word will be " add tiles start menu" .

3.4 Semantic keyset

Based on pre- processed word, a semantic keyset is generated. That is various aspects in which that particular word is used.The semantic keyset is generated by Wordnet tool.The semantic keyset represents the different meaning of a particular word. The similar keyword is clustered into

three types: hypernym, hyponym and synonym. In which hypernym represents the technical meaning of a word, hyponym represents specific meaning of a word and synonym represents the general meaning of a word. This tool is locally available in the web and it is only an additional feature to our proposed system. The generation of these keywords will help the user to make the search more specific. For example, on typing a query " how to make coffee" , the pre- processed keyword will be " make coffee" , the words like bean, drink, etc., will come under the category synonym, the words like espresso, latte, etc., will come under the category hypernym likewise some specific meaning of coffee will come under hyponym.If the user needs search related to any one of the words provided he/ she can simply select the word and continue the search or else directly click the pre- processed query word.

3.5 Search in web server

Then the selected word is given as input to server. The web results given by the server is the webpages, the news and the images for the news. The server search for the web results related to the search. This process is same as that carried out in any web browser.

3.6 Converting the webpages into documents

These webpages are dynamically given into the python code which helps in retrieving the web content. The webpages has to be processed and the web content has to be retrieved for the weight calculation of keyword. The web content can be retrieved in the webpages using a python code embedded in a jsp program. The web content of each web page is stored separately in individual documents correspondingly.

3.7 Documents are processed using WEGA

The web content retrieved from the webpages is subjected to weight calculation of keyword. The WEGA is applied to the calculated weight of the keyword for each document. If a particular query is pre- processed, it may contain one or more strings and they are considered as patterns. Then the particular pattern is searched in all the documents.

Algorithm 1: *Weight- based webpage retrieval algorithm*

Input:
D is the no. of documents.
X is the keyword to be searched.
W is the weight of the word in the document.
S is the no. of sentence in the document.
C is the words in each sentence.
P is the pattern of words

Output:
Documents in the decreased weight order.

Initialization:
Weight of the word to be searched is set to 0

Execution:

```
for i from 1 to  D do
for j  from 1 to S do
for k from 1 to P do
if  Document Di contains the pattern P then
        if  SentencesSj  contains that word C then
       Increment the count
else
skip to the next sentence
end if
else
skip to next document
end if
end for
end for
if weight is less than the threshold value
then
eliminate the document Di
else
compare documents using WEGA algorithm
display the documents in descending order of
weight
then redirect the documents  to the Server to find the
corresponding webpages
end for
```

The documents are taken as input and read one by one. For each document particular pattern is searched. If the document contains that pattern then weightage of that particular word is calculated and weightage is not calculated for that remaining documents. These documents are given as input to selection process of weight enhanced genetic algorithm as shown in table 1.The algorithm for

processing the documents and selecting the best set of documents is described below.

Table 1. Documents and their weights

Documents	Weights	Documents	Weights
Doc 1	25	Doc 4	0
Doc 2	7	Doc 5	0
Doc 3	0	Doc 6	12

Algorithm 2: *Comparing documents using WEGA Algorithm*

Input:
set of documents along with their weights

Output:
Webpages obtained after applying WEGA

Initialization:
The initial set of population

Execution:
Fitness value is taken as 1
if the individuals satisfy the fitness condition
enter it into new population
selecting two individuals for crossover based on fitness condition
applying one point crossover by interchanging the bits
generating child off springs as a result of crossover operation
else
eliminate that individual which does not satisfy the fitness condition
according to the weights of child, the documents are arranged in descending order

For eg. If the given query is **" how to make coffee"** then the pattern to be searched will be **" make coffee" .** This pattern is searched in each document separately. If the document contains that pattern, then the weightage of the word **coffee** is calculated by adding its count dynamically and the weights of each document is stored separately.

3.8 Process in Weight Enhanced Genetic Algorithm

The various process of WEGA involves selection, crossover and mutation. From the below fig 3.2, it describes the process that is used in WEGA. Initially documents and their weights are taken as input and it processed using weightage and final documents are given in descending order for further processing.

a) Selection

These documents and their weights are considered as initial population and they are checked for its Fitness. For example. Fitness value is randomly taken as 1.One by one the weightage of these documents are compared with fitness value.If it satisfies that condition, then solution set is found. Else by applying mutation, bits are flipped slightly. Then again adding the individuals to new population and re- evaluating the individuals against the fitness condition until the best set of individuals are obtained which is shown in table 2.

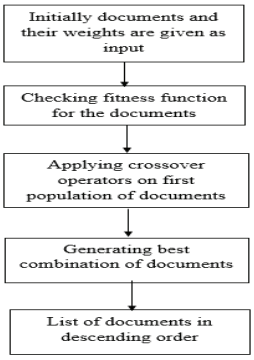

Fig 2. Process of WEGA

Table 2. Documents satisfying their fitness condition

Parents	Documents	Weights
P1	Doc 1	25
P2	Doc 2	7
P3	Doc 3	12

b) Crossover

Now select the two individuals from the new set of population selected. These two individuals are subjected to crossover operation. The crossover operation here done is a simple one point crossover. The weights of the documents are converted into bit string as shown in table 3.

Table 3. Weights are converted into bit string

Parents	Documents	Bit String
P1	Doc 1	11001
P2	Doc 2	01010
P3	Doc 3	11111

The crossover operation is done by interchanging the bits of two parents slightly and creating the children as shown in table 4.For example, let us take two parents s1 and s2 where s1=000000 and s2=111111. On applying crossover operation, the off springs s1' and s2' are produced which is s1' =110000 and s2' =001111. These two off springs are created using one point crossover operation and are put into next generation. By recombining the portion of good individuals, this process is more likely to produce even better individuals.

Table 4. Crossover of two parents

Parents	Bit String	Childs	Bit String	Weight
P1	11| 001	C1	11111	31
P2	01| 111	C2	01001	9

c) Mutation

If certain individual does not satisfy the fitness condition, those individuals are subjected to the mutation operation. The mutation operation is done by flipping the bits of the individual slightly as shown in fig 3.After mutation is carried out, the mutated individuals are added into the new population. The new population is re- evaluated against the fitness condition. This process continues until the best set of individuals are selected.

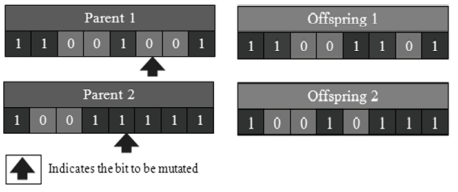

Fig 3. Mutation operation

The above process of WEGA is applied until the fittest webpages to be displayed to the user is identified. Only the fittest webpages will be displayed to the user. The fittest webpages here represents the webpages which are more relevant to the search whereas the webpages which are less relevant are undergone mutation process, then again the fitness condition is checked. If again the webpages are less relevant, those webpages are eliminated. Then Final documents are redirected to the server and the web pages corresponding to the documents are identified. The final optimized webpages are displayed to the user.

4. Result Analysis

In this chapter, the performance of proposed algorithm is compared with other traditional approaches. The performance of various algorithm is analysed based on the parameters such as Accuracy, Memory Used and Time Consumption. The memory used to perform a particular search should be less, accuracy should be more and time consumption should be less compared to other algorithms.

4.1 Memory Used

The memory consumption shows the amount of main memory required to process the algorithm task. That is also known as the space complexity of algorithm. Graph represents the memory consumption to process the task. Where X axis of graph shows the number of searches and the Y axis shows memory consumption in kilobytes. According to the experimented results the amount of memory consumption is similar and not more fluctuating. But the respective proposed genetic approach is more efficient than other traditional approaches as shown in fig 4.

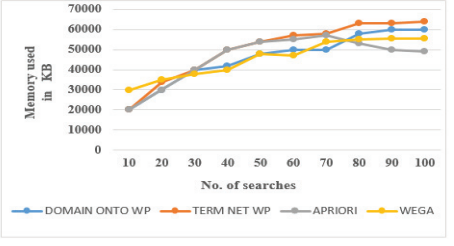

Fig 4. Memory used to perform the task

4.2 Accuracy

The accuracy of the proposed algorithm provides the amount of generated recommendation is better

than the actual outcomes by other models. It defines the amount of precise webpage recommendation by taking weightage of the keyword as parameter. Fig 5 represents the graph showing percentage of accurate recommendation of web pages where X axis of graph shows the no. of searches and the Y axis shows accuracy in percentage.

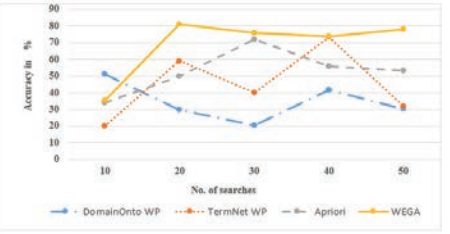

Fig 5. Accuracy of webpages recommended

4.3 Time Consumption

The amount of time required to provide exact webpages as per user' s request is termed here as the time consumption. According to observations the amount of time taken by proposed genetic system for processing the request is not much fluctuated and not also affected by the amount of data to be process. The comparative results of the systems show the effectiveness of the proposed technique that consumes less time to process the query and to display the webpages to user as compared to traditional approach.

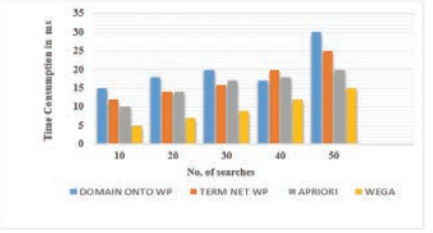

Fig 6: Time consumed by the system

Fig 6 represents graph showing the recommendation time taken by proposed system to generate recommendations. Where X axis of graph shows the number of searches and the Y axis show time consumed in milliseconds.

5. Conclusion And Future Work

The proposed system is mainly concerned about providing the accurate set of webpages relevant to the search. For providing the user with the optimized webpages it uses the WEGA. The weightage is calculated for the web contents retrieved from the webpages given as a output for the search query by Google API. The WEGA is then applied to the weights calculated for the document. The weights of the documents which are selected as initial population are checked against the fitness condition.

The eligible webpages are subjected to the crossover operation and off springs are generated. The weights of the offspring are linked to the parent id. Now the documents are arranged in the descending order of weight. Then the documents which satisfies the fitness conditions are redirected to the server and the corresponding webpages are displayed to the user. It is more advantageous than the other system as it does not display the webpages to the user based on the navigational pattern rather it concerns on the exactness of search. The proposed system is evaluated based on various parameter. The proposed system proves to show high performance compared to other algorithms and traditional methods.

The future work we suggest to this project is described as follows. In the proposed system the weightage is calculated based on the number of times the keyword is repeated in the document. As a future work, not only the weightage of keyword is taken and also the weightage of the meaning of the keyword is also taken into account during the weight calculation. This will help the search to be further more specific.

References

1. Chintankumar S. Maisuriya, Mr.Vaibhav Gandhi," A Review on User" s Future Request Prediction in Web Usage Mining" , International Journal of Advanced Research in Computer Science and Software Engineering,Volume 4, Issue 12, December 2014.
2. Philomina Simon ," Two Stage Approach to Document Retrieval using Genetic Algorithm" , International Journal of Recent Trends in Engineering, Vol. 1, No. 1, May 2009.
3. Mr. A. Kalayanasaravan, Dr. M. Thangamani, Dr. E. T. Venkatesh," Document Retrieval System using Genetic Algorithm" , International Journal of Scientific Engineering and Technology Volume No.2, Issue No.10, pp : 943- 946.

4. SheetalKumrawat, Pramod S. Nair ," Web Page Recommendation Using Efficient Weight Based Prediction System" , International Journal of Science and Research (IJSR),November 2015.
5. RakhiChakraborty ," Domain Keyword Extraction Technique: A New Weighting Method Based On Frequency Analysis" ,Computer Science & Information Technology (CS & IT), pp. 109–118, 2013 .
6. Modraj Bhavsar,Mrs. P. M. Chavan," Web Page Recommendation Using Web Mining" , Modraj Bhavsar Int. Journal of Engineering Research and Applications Vol. 4, Issue 7(Version 2), July 2014, pp.201- 206.
7. Prince Mary.Sand E. Baburaj," Constraint Informative Rules for Genetic Algorithm- Based Web Page Recommendation System" , Journal of Computer Science 9 (11): 1589- 1601, 2013.
8. GunjanVerma,VineetaVerma," Role and Applications of Genetic Algorithm in Data Mining" ,international Journal of Computer Applications (0975 – 888) Volume 48– No.17, June 2012.
9. ThiThanh Sang Nguyen, Hai Yan Lu, and Jie Lu," Web- Page Recommendation Based on Web Usage and Domain Knowledge" , IEEE Transactions on Knowledge and Data Engineering, vol. 26, no. 10, october 2014.
10. Sangari Devi, Dr.S.Dhinakaran," Crossover and Mutation Operations in GA- Genetic Algorithm" , International Journal of Computer & Organization Trends –Volume3 Issue4 – May 2013.
11. B. Liu, B. Mobasher, and O. Nasraoui, " Web usage mining," in Web Data Mining: Exploring Hyperlinks, Contents, and Usage Data, B. Liu, Ed. Berlin, Germany: Springer- Verlag, 2011, pp. 527–603.